SENSEM: base de datos verbal del español[1]

Ana Fernández[1], Gloria Vázquez[2] e Irene Castellón[3]

[1]Dep. Fil. Inglesa y Germ., Universitat Autònoma de Barcelona (EUIS),
Emprius, 2, 08202 Sabadell – ana.fernandez@uab.es
[2]Dep. Inglés y Lingüística, Universitat de Lleida,
Pl. Víctor Siurana, 1, 25003 Lleida – gvazquez@dal.udl.es
[3]Dep. Lingüística, Universitat de Barcelona,
Gran Via Corts Catalanes, 585, 08007 Barcelona – icastellon@ub.edu

Resumen. En este trabajo presentamos una base de datos léxica de los verbos del español en la que se ha descrito el comportamiento sintáctico-semántico de aproximadamente unos 1.100 predicados de esta lengua. Los términos de descripción incluyen papeles semánticos, estructura argumental, semántica oracional y preposiciones. Uno de los aspectos más importantes de dicho recurso es el tratamiento de la interfaz sintáctico-semántica, ya que para cada verbo se indica la lista de esquemas sintácticos que acepta asociados a su significado oracional.

Summary. In this paper we present a lexical database of Spanish verbs that describes the syntactic-semantic behaviour for approximately 1,100 predicates in this language. The terms of description include semantic roles, argument structure, oral semantics and prepositions. One of the most important features of this resource is the manner in which it treats the syntactic-semantic interface, indicating a list of syntactic frames and the sentential semantics for each of them.

Palabras Clave: BD verbal, esquemas sintáctico-semánticos, papel semántico.

1. Introducción

La base de datos SENSEM contiene información sobre verbos del español y sobre su comportamiento en la oración[2]. La unidad de descripción es el sentido, es decir, un mismo verbo podrá tener diferentes descripciones, una por cada sentido observado.

En la actualidad, la BD incluye un total de 1.092 sentidos derivados de 788 formas verbales diferentes. La diferenciación de sentidos es propia de nuestro trabajo aunque se ha partido de diferentes fuentes diccionariales.

Para cada sentido se han descrito las relaciones semánticas que se establecen con sus argumentos en términos de papeles semánticos. Los papeles semánticos están directamente relacionados con una de las posibles estructuras sintácticas en las que puede aparecer cada verbo y que se especifican en las entradas léxicas.

[1] Este artículo se ha realizado mediante la ayuda del MCyT (BFF2003-06456) al proyecto "Creación de una base de datos de semántica oracional y un entorno de consulta y gestión".

[2] Esta base de datos es una evolución de la base de datos confeccionada para VOLEM (ref. ABM/acs/XTI-CTP 2000-1).

Las estructuras sintácticas se expresan a través del número y del tipo de sintagmas que el verbo subcategoriza. También se expresan las preposiciones para cada SP. Además, cada estructura de subcategorización se presenta asociada a una interpretación semántica (causatividad, anticausatividad, etc.), ya que una de las hipótesis teóricas de nuestro modelo descriptivo es que las estructuras sintácticas expresan un significado propio independiente de la pieza léxica verbal que se incluye en dichas estructuras y con la que, composicionalmente, las oraciones adquieren su significado pleno. En teoría lingüística ésta no es una idea nueva sino que ha sido defendida por diversos autores dentro de lo que se llama la gramática de las construcciones (Fillmore 1985, 1988, Goldberg 1994). Asimismo, nuestra descripción del léxico también se basa en las ideas de las teorías más lexicalistas que postulan la organización de las construcciones desde un enfoque más léxico en las que se considera que los componentes semánticos de los predicados son los que determinan que un verbo participe en una diátesis determinada (Levin 1993, Pinker 1989).

El resultado es la constitución de una base de datos de un conjunto significativo de los verbos del español que contiene información a dos niveles: por un lado, a nivel oracional se explicita la semántica y a nivel argumental los diferentes constituyentes se caracterizan a través de sus propiedades sintáctico-semánticas. El listado de interpretaciones semánticas para el español que se ha confeccionado hasta el momento es de 11. De las posibles combinaciones entre estas etiquetas y los esquemas sintácticos se obtiene un total de 76.

Toda esta información está contenida en una única tabla de datos que se encuentra disponible en dos formatos, Access y SQL. Los campos que se incluyen en dicha tabla son los siguientes:

- Identificación del sentido
- Definición del sentido
- Ejemplos de uso
- Esquemas sintáctico-semánticos
- Lista de papeles semánticos
- Preposiciones

Creemos que la creación de este recurso supone una aportación de gran utilidad, dado que hay muy pocos léxicos de este tipo para el español (García de Miguel *et al*. (prensa), Rojo 2001, Subirats y Petruck 2003), que sí existen para el inglés. Además, la inclusión de la semántica oracional supone un valor añadido en tanto que dota a la descripción realizada de un nivel de información más que consideramos fundamental en el tratamiento de la interfaz sintáctico-semántica de los verbos. Este nivel de descripción permite tratar los fenómenos estructurales desde un punto de vista semántico, lo cual presenta muchas ventajas tanto a nivel intralingüístico como interlingüístico.

En los apartados siguientes vamos a explicar con más detalle la metodología utilizada en el diseño y composición del recurso así como el tipo de representación escogida para la codificación de los datos. Así, dedicamos el apartado 2 a explicar cómo se han identificado los diferentes sentidos de un lema verbal. En el apartado 3 pasamos a describir los esquemas sintáctico-semánticos utilizados. El apartado 4 lo dedicaremos a la presentación de los papeles semánticos y la codificación de las preposiciones. Finalmente, presentaremos las conclusiones y el trabajo en curso.

2. Identificación de sentidos

La primera cuestión metodológica importante que se ha tenido que afrontar es definir qué se considera sentido en nuestra base. La identificación de los diferentes sentidos de un lema se ha llevado a cabo mediante el análisis comparativo de las entradas de varios diccionarios (DRAE, Salamanca, María Moliner, entre otros). Cada sentido verbal se ilustra mediante ejemplos, combinando la introspección con consultas a corpus a través de Internet (corpus de la RAE, www.corpusdelespanol.org) y consultas efectuadas en recuperadores de información (Google).

El primer criterio que se ha tenido en cuenta es el de priorizar los sentidos más frecuentes, dejando de lado los de uso muy restringido u obsoletos, así como los basados en locuciones verbales y auxiliares. Por otro lado, los sentidos se han establecido en función de la existencia de claras diferencias semánticas[3], lo cual, a su vez, tiene normalmente repercusiones importantes en la subcategorización (diferentes tipos de sintagmas, sobre todo SN vs. SP), en las restricciones de selección y en la semántica oracional (Gutiérrez 1989, Cruse 1995, Palmer 2000). Así pues, se ha priorizado en este punto la sintaxis sobre la semántica, evitando asociaciones de tipo cognitivo si había una base estructural para mantener la diferenciación, ya que la finalidad de nuestro recurso es que sirva para el análisis de las construcciones oracionales del español. El resultado obtenido es el de 4 sentidos de media por verbo. Se ha intentado evitar caer en la subespecificación de acepciones y, por tanto, las unidades que hemos definido son mucho más amplias que en otros recursos lexicográficos como WordNet (Miller *et al*. 1993), en los que la granularidad es mucho mayor.

A continuación presentamos un ejemplo en el que se puede observar cómo han incidido las diferencias sintáctico-semánticas a la hora de diferenciar sentidos para el verbo *dejar*, que presenta matices de significado que podrían ser considerados susceptibles de ser agrupados. En el primer sentido, la subcategorización incluye 3 argumentos obligatorios y no se permite ninguna otra estructura además de la que expresa agentividad; en el segundo, en cambio, uno de los argumentos (el que expresa el destinatario) es opcional y es posible tanto utilizar una construcción agentiva como pasiva o recíproca.

- Sentido 1: dar una persona una cosa **en herencia**:
  agentiva- SN V SN SP: El abuelo ha dejado todo lo que tenía a sus nietos
- Sentido 2: poner una persona una cosa que posee en manos de otra **por un tiempo**
  agentiva – SN V SN: Mi hermano nunca deja sus cosas
  agentiva – SN V SN SP: Mi hermano nunca me deja sus cosas
  pasiva – SN SE V: Siempre se dejan los trastos más viejos
  recíproca – SN PRON V SN: Entre hermanos nos lo dejamos todo

[3] En este sentido, nos diferenciamos de otras aproximaciones, como es el caso de la base de datos ADESE, donde se utiliza un concepto más amplio de sentido.

3. Los esquemas sintáctico-semánticos y el papel de la semántica oracional

La definición de los patrones sintáctico-semánticos es la información central de la base de datos que presentamos. Tal y como ya hemos comentado, a diferencia de otras bases léxicas que existen, en la nuestra no nos limitamos a expresar la estructuras de subcategorización o las estructuras sintácticas de los predicados, sino que a cada esquema sintáctico del estilo SN V SN se le asigna una estructura semántica que expresa el significado de dicho esquema y que formará parte del significado de cualquier oración en la que se use ese verbo en una configuración sintáctica como la ejemplificada en nuestra base. Así, la descripción se establece en términos de pares de forma y significado sobre los que se afirma o niega su inclusión. Por ejemplo, para el patrón anteriormente citado, expresamos ahora sus distintos significados oracionales con una primera etiqueta abreviada, que aparece asociada a la configuración sintagmática de dicho patrón, también abreviada[4]:

caus-2np
proc-2np

En el primer caso, se trata de una estructura que requiere dos sintagmas nominales, uno sujeto y otro objeto, que expresa un significado causativo. Dicha semántica, en nuestra aproximación implica que se trata de un evento con un límite incluido, como en el caso de *romper*. En el segundo caso se describe la misma configuración sintáctica con una semántica distinta, ya que se trata de un proceso, es decir, de una actividad sin límite, como por ejemplo *entender*.

Las semánticas que se han definido para etiquetar las estructuras sintácticas son: causativa, agentiva, proceso, antiagentiva, anticausativa, pasiva, impersonal, estativa, reflexiva, recíproca y resultativa.

Por lo que se refiere a los patrones sintácticos, se han usado los siguientes:

np	np-pp	np-2pp
np-adj	np-advp	np-advp-pp
2np	2np-2pp	2np-adjp
2np-advp	2pp 3np	
3pp	pp pp-adjp	

Además de estos patrones sintácticos, también se han tenido en cuenta las siguientes construcciones perifrásticas[5]:

Dejar + adjp-2np	Hacer + inf-2np	Hacer + compl-2np
Dejar + part-2np	Hacer + inf-2np-pp	Estar + adjp-np
Dejar + part-2np-pp	Hacer + compl-2np-pp	

Ejemplos de oraciones que se expresarían con alguno de los patrones anteriores son las siguientes:[6]

---

[4] Usamos las siglas de los sintagmas en inglés y en minúscula. Se prescinde de la etiqueta <v>, que representa el verbo y que se da por supuesto, y se especifica el número de constituyentes del mismo tipo sin repetir la sigla, en este caso, 2np.

[5] Otro patrón sintáctico-semántico que se ha empezado a codificar en la base de datos es el del dativo de interés, aunque todavía no se ha realizado un estudio exhaustivo.

Eventos agentivos:
np: María anduvo toda la noche en busca de refugio
np-pp: María anda por el campo toda la noche en busca de refugio
2np-pp: El martes María envió el paquete a Juan
np-2pp: María habló con su abuela sobre la guerra civil

Eventos anticausativos:
pr-np: La puerta se ha roto
pr-np-pp: María se ha sorprendido con el comentario de Pepe
np: Los datos han variado

Construcciones perifrásticas:
hacer
hacer-inf-2np: Ese comentario hizo enfadar a Juan
dejar
dejar-part-np-pp: Juan ha dejado confundida a María con su comentario

El total de combinaciones posibles entre las etiquetas semánticas de las oraciones y los diferentes esquemas sintácticos con los que se pueden combinar y que se han recogido en este trabajo es de 76.

Dentro de este apartado, los problemas que se han encontrado tienen que ver con la distinción entre adjunto y argumento. Un criterio que parece indiscutible es la obligatoriedad a la hora de identificar un argumento. Pero ello no quiere decir que un constituyente opcional no pueda ser argumento. Así, tenemos en cuenta el alcance de la semántica del verbo y cómo ésta se expresa sintagmáticamente. Por ejemplo, muchos verbos psicológicos que expresan un cambio de estado (*aburrir, alegrar*, etc.) pueden expresar de forma discontinua la causa y uno de los segmentos puede ser opcional, como ocurre con el SP del ejemplo siguiente:

"El discurso del conferenciante me aburre"
"El conferenciante me aburre (con su discurso)"

Por otro lado, los verbos de trayectoria pueden expresar ésta con diversos argumentos:

"Ya ha venido de Madrid (origen)"
"Venía por la calle (ruta)"
"Venía hacia aquí (destino)"

En estos casos, preferimos hablar de un argumento "trayectoria" que puede expresarse con diferentes puntos de ésta, incluso expresarse más de uno a la vez:

"Ha venido [[de Madrid] [hasta Barcelona] [por autopista]]"

No obstante, no faltan los casos fronterizos y problemáticos, como muchos instrumentos y maneras:

"Ha cargado el camión –con la pala–"
"Ha golpeado a María –con el bate de béisbol–"

o algunas cantidades:

"Ha sufrido –mucho– su pérdida"

---

[6] Los sintagmas no subrayados no son considerados argumentales.

Este punto de partida tiene implicaciones en otros niveles ya que todos los elementos sintácticos que se caracterizan como argumentales han de tener una caracterización sintáctica (tipo de sintagma) y también semántica (papel semántico).

4. Los papeles semánticos y las preposiciones

Como ya se ha expresado anteriormente, en la entrada verbal cada posición sintáctica está relacionada con un papel semántico que caracteriza el argumento en cuestión por lo que respecta a su relación con el predicado. Los papeles se asignan en una lista que se relaciona con una de las construcciones declaradas mediante un operador. La distribución del resto de los papeles en las demás construcciones o su participación en una de ellas se establece mediante reglas de *linking* desde las construcciones a las listas.

Los papeles semánticos se entienden en este trabajo en la línea de su uso en autores como Fillmore 1968 y Dowty 1991, pero se ha ampliado el número de etiquetas usadas y se han incluido nuevos conceptos con el fin de poder dar cuenta de la variabilidad semántica existente cuando se realiza la descripción de un léxico a gran escala. En total se han utilizado 29 etiquetas. Entre éstas se han definido 14 etiquetas a un nivel más abstracto de descripción que nos permiten expresar generalizaciones en aquellos casos en que se considere oportuno:

Inic-iniciador	Ter-tema estado resultado	Ma-manera
Exp-experimentador	Tray-trayectoria	Id-identificador
Tg-tema general	Cb-cambio	Sust-sustitución
Th-tema holístico	Fin-finalidad	Tp-temporal
Ti-tema incremental	Instr-instrumento	

El iniciador indica el argumento que inicia la acción y el experimentador se usa para los participantes con acción psicológica. El papel semántico tema ha sido típicamente usado como cajón de sastre para expresar aquellos argumentos que sufren de alguna manera la acción realizada por el iniciador. Como se puede observar, hemos intentado evitar caer en este error considerando como papeles temáticos distintos los temas incrementales, holísticos y de estado resultado. No obstante, no hemos podido descartar el uso de una etiqueta más amplia, tema general, para aquellos casos puntuales en que no ha sido posible especificar más.

Los temas holísticos se usan para dar cuenta de las entidades desplazadas (*X pone Y en Z*), los de estado resultado se utilizan para caracterizar aquellos constituyentes que expresan el tipo de cambio de estado experimentado por una entidad (*X convierte Y en Z*), los temas incrementales describen aquellos argumentos que son afectados por la acción (*X estropea Y*) y, por último, los temas generales quedarían para constituyentes cuya semántica no ha sido claramente definida, como *X basa Y en Z.*

Nos quedan por comentar los papeles finalidad, instrumento y manera, que son habituales en las listas de papeles semánticos, y los de cambio, identificador y sustitución, que son propias de nuestra descripción. El cambio denota un argumento complejo que integra los diferentes puntos y aspectos de un cambio de estado (estado inicial y estado resultante), como en *X varía de Y a Z*; el identificador se usa para indicar una

cualidad de otro argumento (*X considera Y a Z*), y el papel sustitución describe el participante al que substituye el iniciador (*X habla por Y*).

De estas 14 etiquetas generales, se han subespecificado 4 (iniciador, tema incremental, trayectoria y temporal) y se han creado 14 etiquetas específicas, que nos permiten expresar con más detalle las relaciones semánticas de los argumentos cuando ello es posible:

- Iniciador: el que inicia la acción puede ser un experimentador[7], el cual en este caso está asociado típicamente a acciones de tipo mental no causativas[8] (*pensar*), o un agente o una causa. Las diferencias entre agente (*X come Y*) y causa (*X aburre a Y*), sin embargo, no siempre son claras. En principio, el primero pero no el segundo actúa con voluntariedad pero en ambos casos estos participantes provocan claramente la acción. En algunos casos no es relevante esta característica de voluntariedad (*X rompe Y*) y, entonces, se usan ambos papeles. Los casos en que se usa la etiqueta general iniciador son aquellos en los que dicho argumento tiene un papel poco activo en la iniciación, como en el caso de *lograr* o *necesitar*.

- Tema incremental: cuando la afección es negativa, hablamos de tema incremental víctima (*X pega a Y*); si la afección es positiva, se trata de un tema incremental beneficiario (*X regala Y a Z*); si el resultado de la acción es la creación de una nueva entidad, hablamos de tema incremental de creación (*X construye Y*), y si es la destrucción de una entidad ya existente, de tema incremental de destrucción (*X destruye Y*).

- Trayectoria: distinguimos 4 subtipos: la localización, que expresa la situación exacta o aproximada donde tiene lugar la acción, como en *X vagabundea por Y*, el destino (*X va a Y*), el origen (*X viene de Y*) y la ruta (que expresa el total del desplazamiento o una porción, como en *X introduce Y por Z*)

- Temporal: desde la perspectiva del tiempo también podemos expresar el destino (*X acaba a las Y*), el origen (*X empieza a las Y*) o la localización (que da indicaciones sobre el momento en que tiene lugar la acción, *X ocurrió a las Y*).

Para la codificación de las preposiciones, éstas se han clasificado en función del tipo de papeles semánticos con los que se pueden asociar. Así, cuando un argumento preposicional se puede expresar con todas las preposiciones asociadas a su papel semántico, éstas no se especifican, ya que resulta redundante. Las preposiciones sólo se codifican en aquellos casos en que el verbo no permite la aparición de todas ellas sino de un subconjunto de las mismas, que es lo más habitual.

5. Conclusiones y trabajo en curso

En el presente artículo hemos presentado una base de datos que explicita el comportamiento verbal de unos 1.100 verbos de la lengua española por lo que respecta a la interfaz sintáctico-semántica. La base de datos incluye para cada sentido una definición, ejemplos de uso y varios campos que definen la descripción sintáctico-

[7] Se diferencia entre inic_exp (*X piensa*) y exp (*X gusta a Y*) según la posición sintáctica que ocupa en la frase activa, lo cual a su vez tiene repercusiones en el tipo de relación semántica que establecen dichos argumentos con el verbo ya que el foco de la frase es distinto en cada caso.

[8] Es decir, que no provocan un estado resultado.

semántica, que se centra en la especificación de las formas posibles de expresión (sintaxis) de determinados significados oracionales descritos para el español, de los papeles semánticos que caracterizan los argumentos de los verbos en las construcciones y de las preposiciones que requieren.

En la actualidad se está trabajando en dos líneas básicamente. Por un lado, se está llevando a cabo la conexión de cada sentido con la red semántica WordNet. En el momento en que se realizó este artículo se habían establecido 150 relaciones entre sentidos y *synsets*. También estamos utilizando la clasificación semántica verbal de este recurso para agrupar los diferentes verbos en conjuntos nocionales, información que puede ser útil para aplicaciones en las que se trabaje en ámbitos restringidos del lenguaje.

Por otro lado, estamos llevando a cabo paralelamente la anotación de un corpus del español de un millón de palabras siguiendo los criterios presentados en este trabajo. El objetivo es doble: en primer lugar, se pretende comprobar hasta qué punto la descripción teórica de los verbos se realiza en el uso "real" del lenguaje y, en segundo lugar, pretendemos ampliar la información del recurso actual con la codificación de nuevos verbos, elegidos por su frecuencia, y con nuevos datos sobre construcciones que se observen en el corpus analizado.

Bibliografía

Cruse, D. A. (1995) "Polysemy and related phenomena". En *Computational Lexical Semantics* (Saint-Dizier, P. y E. Viegas eds.), Cambridge University Press.

Dowty, D. (1991) "Thematic Proto-Roles and Argument Selection", *Language*, 67(3), 547-619.

Fillmore, C. J. (1968). "The case for case". En *Universals in Linguistics* (E. Bach, R. T. Harms, eds.), Nueva York: Holt, Rinehart, Winston.

Fillmore, C. J. (1985) "Syntactic Intrusions and the Notion of Grammatical Construction", *BLS*, 11, 73-86.

Fillmore, C. J. (1988) "The Mechanisms of "Construction Grammar", *BLS*, 14, 35-55.

García-Miguel, J. M, L. Costas y S. Martínez (en prensa). "Diátesis verbales y esquemas construccionales", *Actas del VI Congreso Internacional de Lingüística Hispánica*, Leipzig, 2003.

Goldberg, A. E. (1994) *Constructions: A Construction Grammar Approach to Argument Structure*, Chicago, Illinois: University Chicago Press.

Gutiérrez, S. (1989) *Introducción a la semántica funcional*. Madrid: Síntesis

Levin, B. (1993) *English Verb Classes and Alternations*, Chicago: The University of Chicago Press.

Miller, G. A., R. Beckwith, Ch. Fellbaum, D. Gross y K. Miller (1993) "Introduction to WordNet: An On-line Lexical Database", *Journal of Lexicography*, 3(4), 234-244.

Palmer, M. (2000) "Consistent criteria for sense distinctions", *Computers and the Humanities*, 34, 217-222.

Pinker, S. (1989) *Learnability and cognition: The acquisition of argument structure*, MIT Press

Rojo, G. (2001). "La explotación de la Base de Datos Sintácticos del español actual". En *Gramática española. Enseñanza e investigación* (J. Kock, ed.), I, 7. Universidad de Salamanca.

Subirats-Rüggeberg, C. y M. R. L. Petruck (2003) "Surprise: Spanish FrameNet! Presentation at Workshop on Frame Semantics", *Proceedings of the International Congress of Linguists*, Praga.